EM algorithm

The example in the book for doing the EM algorithm is rather difficult, and was not available in software at the time that the authors wrote the book, but they implemented a SAS macro to implement it. Variations on this EM algorithm have since resulted in research papers, some of which are still fairlly recent.

Rather than jumping into the details of the particular EM algorithm, we'll look at a simpler example to get the idea of how it works

EM algorithm

We'll do an example with coin flipping. You are given two coins, A and B, where the probability of heads for each coin is θ_A and θ_B, respectively. You wish to estimate $\theta = (\theta_A, \theta_B)$.

In the first part of the experiment, you randomly select a coin, then flip it 10 times and keep track of the number of heads. You repeat this procedure five times, so there are a total of 50 coin flips. You could represent the data by $(x_1, \ldots, x_5)$ and $(z_1, \ldots, z_5)$ where x_i is the number of heads on the ith set of 10 flips, and z_i is an indicator of which coin was chosen.

EM algorithm

Assuming that z_1 is not constant (you picked both coins at least once), then you can estimate θ by

$$\widehat{\theta}_A = \frac{\text{number of heads with } A}{\text{number of flips with } A}$$

$$\widehat{\theta}_B = \frac{\text{number of heads with } B}{\text{number of flips with } B}$$

This is the intuitive estimator and is also the maximum likelihood estimator.

EM algorithm

Equivalently, we could write

$$\widehat{\theta}_A = \frac{\sum_{i=1}^{5} x_i I(z_i = A)}{10 \cdot \sum_{i=1}^{5} I(z_i = A)}$$

$$\widehat{\theta}_B = \frac{\sum_{i=1}^{5} x_i I(z_i = B)}{10 \cdot \sum_{i=1}^{5} I(z_i = B)}$$

Assuming that the denominators are non-zero.

EM algorithm

Where things get more difficult is if you randomly pick a coin each time but you don't know which coin you picked. What you observe are the counts $x = (x_1, \ldots, x_5)$, but you don't observe $z = (z_1, \ldots, z_5)$. If the coins have very different biases, such as $\theta_A = 0.95$ and $\theta_B = 0.1$, then you would have a good guess at what the z values are.

If you observe $x = (9, 2, 10, 0, 0)$, then you could feel fairly confident that you picked coin A on the first and third trials, and coin B on the second, fourth, and fifth trials. However, if $\theta = (0.5, 0.6)$, then it will be very difficult to tell which trials were from which coin.

EM algorithm

For the case where z is unobserved, the variables in z are often called hidden variables or latent variables. (This is analogous to the random coefficients in the frailty model.

This is where the EM algorithm comes in. A simplification of the algorithm is the following. Start with an initial guess, $\theta^{(0)} = (\theta_A^{(0)}, \theta_B^{(0)})$. Given this initial guess for θ, you could then take a best guess for $z = (z_1, \ldots, z_5)$. Given this value for z, can then get a likelihood estimate for the data. The likelihood of the data is

$$L(\theta_A, \theta_B) = \prod_{i=1}^{5} \binom{10}{x_i} \theta_i^{x_i} (1 - \theta_i)^{10 - x_i}$$

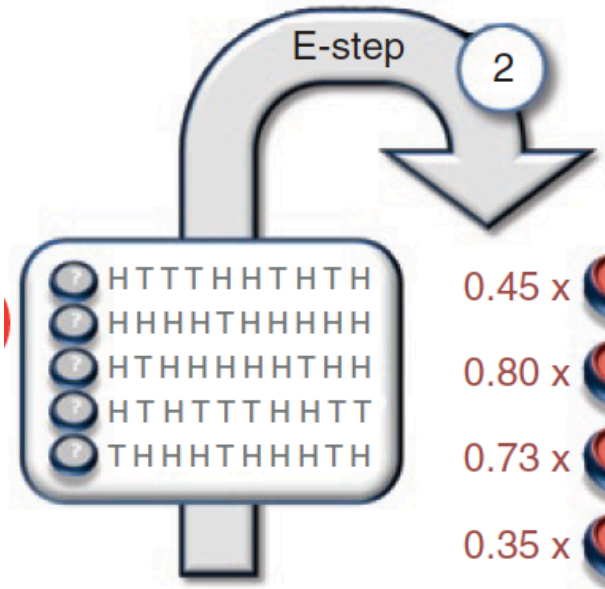
where $\theta_i = \theta_A$ if $z_i = A$ and $\theta_i = \theta_B$ if $z_i = B$. Given the likelihood function, you get an updated estimate of θ which you can call $\theta^{(1)} = (\theta_A^{(1)}, \theta_B^{(1)})$.

EM algorithm

Now that you have an updated estimate for θ, this will result in a possibly new best guess for z, which result in a new estimate of the likelihood, and so on. If the best guess for z doesn't change from one iteration to the next, then the algorithm stops.

For the actual EM algorithm, instead of using the best guess for z, we instead use the probability that the coin picked was A versus B given the data and our current guess of θ. This requires Bayes theorem, since we get $P(A|\text{data})$ from $P(\text{data}|A)$.

EM example



Coin A	Coin B
0.45 x A	0.55 x B
0.80 x A	0.20 x B
0.73 x A	0.27 x B
0.35 x A	0.65 x B
0.65 x A	0.35x B

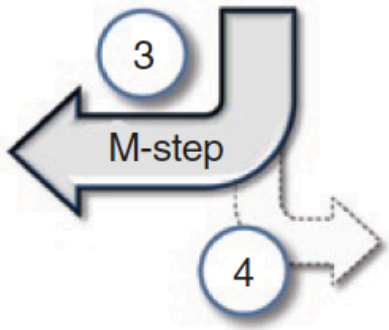
Coin A	Coin B
≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$\hat{\theta}_A^{(0)} = 0.60$

$\hat{\theta}_B^{(0)} = 0.50$

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$



$\hat{\theta}_A^{(10)} \approx 0.80$

$\hat{\theta}_B^{(10)} \approx 0.52$

EM example

Here we use Bayes Theorem for each of the five samples to estimate the probabilities that each set of 10 coin flips came from that particular coin. This gives a distribution on the possible coins that were flipped. In reality only one coin was flipped for each set of 10, but we use a distribution to reflect our uncertainty on which coin was flipped.

For example, the probability that the first coin flipped was A given that 5 out of 10 flips were heads and that our initial guess is $\theta_A = 0.6$ and $\theta_B = 0.5$ is roughly 0.45, and $P(B|5) = 0.55$.

EM example

An easier way of doing this is with the idea of membership weights. Here, let the 5 heads in the first toss belong to A with weight $5(.45) = 2.25$ and the 5 heads belong to B with weight $5(.55) = 2.75$. There are four membership weights which add up to 10 for each observation: heads belonging to A, heads belonging to B, tails belonging to A, tails belonging to B.

Then we update $\widehat{\theta}_A^{(0)}$ using the probability of belonging to A given heads, which is

$$\theta_A^{(1)} = \frac{21.3}{21.3 + 8.6} = 0.71$$

and

$$\theta_B^{(1)} = \frac{11.7}{11.7 + 8.4} = 0.58$$

The 10th iteration should be

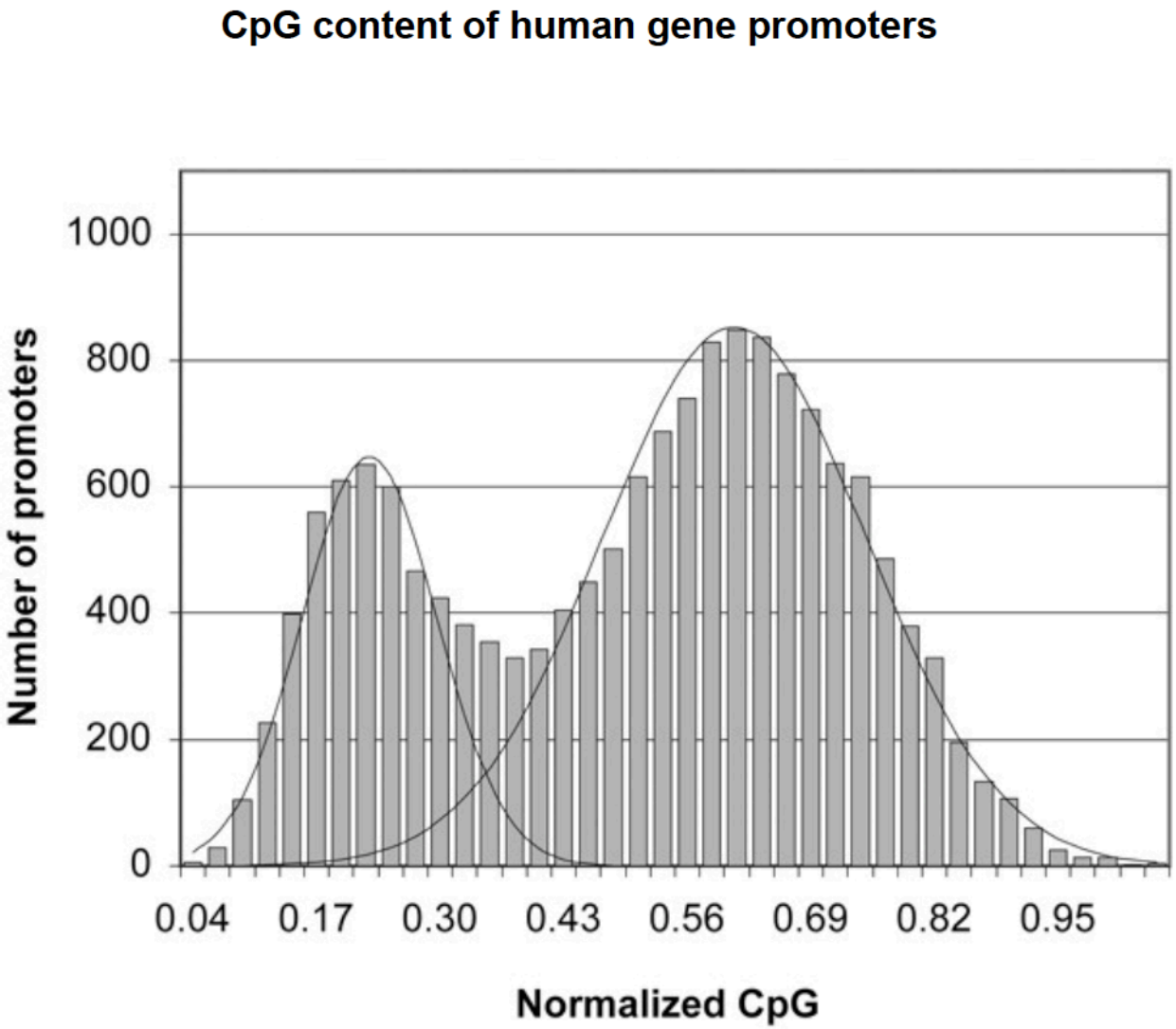
$$\theta_A^{(10)} = 0.80, \quad \theta_B^{(10)} = 0.52$$

CpG example

Another example which is more common is to use the EM algorithm to estimate the mixture parameter in a mixture of normals (or other distributions). In this example, DNA sequences are sampled from different parts of the human genome. One statistic often extracted from DNA sequences is the proportion of the sequence that has the letter C followed by G, often notated CpG. Approximately 42% of the human genome has C or G as the DNA letter, and the expected frequency of the dinucleotide CpG is about 4% assuming letters are distributed at random. However, the observed frequency is closer to 1%.

However, genes that have promoter regions that have relatively high frequencies of CpG tend to be expressed more often — that is, the proteins associated with these genes tend to get produced by cells more often than genes that have lower levels of CpG in their promoter regions, so the CpG frequnecy is particularly interesting to look at.

CpG example


CpG content of human gene promoters
Number of promoters
1000
800
600
400
200
0
0.04
0.17
0.30
0.43
0.56
0.69
0.82
0.95
Normalized CpG

Mixture of normals

From the histogram, there is a bimodal distribution of the frequency of CpG, suggesting that there are at least two types of promoters, some with relatively high CpG frequencies, and some with relative low CpG frequencies. This is just used an example of a bimodal distribution that appears to be a mixture of two normal distributions.

Of interest would be to estimate what the probability is that a randomly sampled promoter comes from one of the two classes of CpG frequencies, in other words what is the mixture of the two densities? We can think of there being two normal distributions, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, and for a randomly selected promoter, the density for the frequency of CpG dinucleotides is

$$f(x|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha) = \alpha N(\mu_1, \sigma_1^2) + (1 - \alpha) N(\mu_2, \sigma_2^2)$$

Mixture of normals

Note that this is different from a bivariate normal, which is a distribution which also has 5 parameters (two means, two variances, and a correlation). This is still a univariate distribution. The idea is for each observation X_i, there is a probability α that X_i comes from a $N(\mu_1, \sigma_1^2)$, and a probability $1 - \alpha$ that X_i comes from a $N(\mu_2, \sigma_2^2)$. This is also different from a linear combination of normal random variables, which would only be normally distributed and not bimodal.

A particularly nast case of a mixture of normals is when $\mu_1 = \mu_2$. In this case, the mixture is not bimodal, so it is difficult to detect that there are really distributions at work. Generally, mixtures of normals can also be extended to more than two variables, and then you have mixture components α_i, with $\sum \alpha_i = 1$, multiple means, μ_i and multiple variances σ_i^2, and the probability is α_i that an observation comes from the ith mixture component.

Mixture of normals

We'll restrict our attention to a mixture of two normals. If we knew which population each observation came from, it would be easy to write the likelihood. It would be a product of all of the univariate normal densities for each observation assuming the observations are independent. The difficulty is that for a given observation, you often don't know which normal distribution it came from.

Mixture of normals

Similar to the previous EM example, we can let $(z_1, \ldots, z_n)$ be a vector indicating which normal distribution each observation came from. Let $z_i = 1$ if the observation is from the first density, for example, and otherwise $z_i = 0$. It would be straightforward to write the likelihood if the zs were observed, but they are not observed, so we can't directly put in numbers (or specific densities) for the likelihood. If we knew the zs, we could write the likelihood as

$$\prod_{i=1}^{n} f_i(x_i)$$

as usual, but the problem is that for each i, $f_i = f_1$ or f_2, and we don't know which. We could also write the likelihood as

$$\prod_{i=1}^{n} \sum_{i=1}^{2} \tau_i f_i(x_i)$$

where $\tau_2 = 1 - \tau_1$, but this likelihood, as a product of sums, is intractable, and requires numerical solutions. Since there are 5 parameters, it is a moderatly high dimensional problem to solve numerically.

Mixture of normals

Sometimes people think of this as a chicken and the egg problem. If we knew the zs, we could estimate θ (the set of parameters). And if we knew θ, we could reasonable estimates of the zs (which density each observation came from). Instead of being able to solve either problem, we guess θ, then estimate $\mathbf{z}$. Then use this estimated $\mathbf{z}$ to get an updated estimate of θ, and so on until the process converges, we hope.

The expected value of z_i is the probability that the ith observation belongs to group 1. We get this from Bayes' Theorem

$$E(z_i|x_i) = P(1|x_i) = \frac{P(x_i|1)P(1)}{P(x_i)} = \frac{P(x_i|1)}{P(x_i|1)P(1) + P(x_i|2)P(2)}$$

$$E(1 - z_i|x_i) = P(2|x_i) = ???$$

Mixture of normals: E-step

$$E(1 - z_i|x_i) = P(2|x_i) = 1 - P(1|x_i)$$

Why? Because $P(A|B) + P(A^c|B) = 1$ or

$$P(A^c|B) = 1 - P(A|B)$$

Bayes rule is applied to each observation to get a probability that that observation belongs to group 1. This is the Expectation-step.

Mixture of normals

We can express the likelihood for the ith observation as

$$L(x_i, z_i|\theta) = z_i\alpha_1 f_1(x_i|\theta) + (1 - z_i)\alpha_2 f_2(x_i|\alpha)$$

or

$$L(x_i, z_i|\theta) = [\alpha_1 f_1(x_i|\theta)]^{z_i} \cdot [\alpha_2 f_2(x_i|\theta)]^{1-z_i}$$

Mixture of normals

For the Maximization step, we update the estimated parameters

$$\mu_1^{(1)} = \frac{\sum_{i=1}^{n} P(1|x_i)x_i}{\sum_{i=1}^{n} P(1|x_i)}$$

$$\mu_2^{(1)} = \frac{\sum_{i=1}^{n} P(2|x_i)x_i}{\sum_{i=1}^{n} P(2|x_i)}$$

$$\sigma_1^{2(1)} = \frac{\sum_{i=1}^{n} P(1|x_i)\left(x_i - \mu_1^{(1)}\right)^2}{\sum_{i=1}^{n} P(1|x_i)}$$

$$\sigma_2^{2(1)} = \frac{\sum_{i=1}^{n} P(2|x_i)\left(x_i - \mu_2^{(1)}\right)^2}{\sum_{i=1}^{n} P(2|x_i)}$$

$$\alpha^{(1)} = \frac{1}{n}\sum_{i=1}^{n} P(1|x_i)$$

Mixture of normals: M-step

The convergence is relatively slow compared to some other iterative procedures, such as Newton-Raphson tends to be. It can take 10 or more iterations for a typical example to converge for the mixing parameter to within 0.01 of the correct answer.

We can think of the Maximization step as maximizing the expected value of the log-likelihood, where instead of the full likehood (which depends on both $\mathbf{x}$ and $\mathbf{z}$), we replace $\mathbf{z}$ with $E[\mathbf{z}|\mathbf{x}, \widehat{\theta}]$, where $\widehat{\theta}$ is our current best guess for θ.

Gene counting with ABO blood system

As another example, we'll consider estimating the allele frequencies of the A, B, and O alleles given observed phenotypes, type A, B, AB, and O. Here type A occurs if you have either AO or AA allele combinations, type B occurs if you have either BO or BB alleles, type O occurs only if you have OO, and type AB occurs only if you have AB.

We observe phenotypes A, B, O, AB, so we don't directly observe the genotypes or individual alleles. If we knew the individual genotypes, then it would be straightforward to get the allele counts. For example, if there were n_{AO} people with genotype AO, n_{AA} people with genotype AA, and n_{AB} people with genotype AB, then there would be $n_{AB} + n_{AO} + 2n_{AA}$ copies of A in the sample.

Gene counting with the ABO blood system

Let the unknown allele frequencies be μ_A, μ_B and μ_O, with $\sum \mu_i = 1$. The expected frequencies of the genotypes (assuming the alleles occur randomly) is

$$P(AA) = \mu_A^2 P(AO) = 2\mu_A\mu_O$$
$$P(BB) = \mu_B^2 P(BO) = 2\mu_B\mu_O$$
$$P(AB) = 2\mu_A\mu_B P(OO) = \mu_O^2$$

These are our parameters, and the genotype frequencies are the hidden information.

Gene counting with the ABO blood system

Using Bayes rule, we can compute the probability that someone has genotype AA given that they have blood type A. This is

$$P(AA|\mathrm{A}) = \frac{P(\mathrm{A}|AA)P(AA)}{P(\mathrm{A}|AA)P(AA) + P(\mathrm{A}|AO)P(AO)} = \frac{\mu_A^2}{\mu_A^2 + 2\mu_A\mu_O}$$

Similarly

$$P(AO|\mathrm{A}) = \frac{2\mu_A\mu_O}{\mu_A^2 + 2\mu_A\mu_O}$$

The expected number of individuals with genotype AA is

$$n_{AA}^{(1)} = n_A \times \frac{\mu_A^2}{\mu_A^2 + 2\mu_A\mu_O}$$

And the expected number of individuals with genotype AO is

$$n_{AO}^{(1)} = n_A \times \frac{2\mu_A\mu_O}{\mu_A^2 + 2\mu_A\mu_O}$$

Gene counting with the ABO blood system

Similarly, we can get expected numbers of individuals with any genotype. The numbers of individuals with genotype AB and OO are given directly from the data. Our updated allele frequencies can be obtained once we have these estimates of the genotype frequencies. We get

$$\mu_A^{(1)} = \frac{2n_{AA}^{(1)} + n_{AB}^{(1)} + n_{AO}^{(1)}}{2N}$$

where N is the number of individuals in the sample, and $2N$ is the number of alleles in the sample. We can go back and forth between updating genotype frequencies (the E-step) and allele frequencies (M-step).

Gene counting with the ABO blood system

To give an example, suppose we observe a sample of individuals and get their blood types, with the genotypes unknown:

$$n_{\mathrm{A}} = 186, \quad n_{\mathrm{B}} = 38, \quad n_{\mathrm{AB}} = 13, \quad n_{\mathrm{O}} = 284$$

Suppose our initial guess is that $\mu_A^{(0)} = 0.3$, $\mu_B^{(0)} = 0.2$, and $\mu_O^{(0)} = 0.5$. Then in the first iteration, we get that our guess for the numbers of genotypes AA and AO are

$$n_{AA}^{(1)} = n_{\mathrm{A}} P(AA|\mathrm{A}) = \frac{\mu_A^2}{\mu_A^2 + 2\mu_A\mu_O} = (186)\frac{0.3^2}{0.3^2 + 2(0.3)(0.5)} = (186)(0.23) = 42.9$$

$$n_{AO}^{(1)} = 186 \times \frac{2(0.3)(0.5)}{0.3^2 + 2(0.3)(0.5)} = (186)(0.7692) = 143.08$$

$$n_{BB}^{(1)} = 38 \times \frac{0.2^2}{0.2^2 + 2(0.2)(0.5)} = 6.333$$

$$n_{BO}^{(1)} = 38 \times \frac{2(0.2)(0.5)}{0.2^2 + 2(0.2)(0.5)} = 31.667$$

Gene counting with ABO blood system

Once we have these expected genotype counts, we get the allele counts as

$$\mu_A^{(1)} = \frac{2n_{AA}^{(1)} + n_{AO}^{(1)} + n_{AB}}{2n} = (42.92 \times 2 + 143.08 + 13)/1042 = 0.2321$$

$$\mu_B^{(1)} = \frac{2n_{BB}^{(1)} + n_{BO}^{(1)} + n_{AB}}{2n} = (6.333 \times 2 + 31.667 + 13)/1042 = 0.0550$$

$$\mu_O^{(1)} = \frac{2n_{OO} + n_{AO}^{(1)} + n_{BO}^{(1)}}{2n} = (284 \times 2 + 143.08 + 31.667)/1042 = 0.7129$$

Gene counting with the ABO blood system

Compared to our initial guess, the frequencies of A and B have gone down, and the frequency of O has increased. We reiterate the process, for example using

$$n_{AA}^{(2)} = 186 \times \frac{0.2321^2}{0.2321^2 + 2(0.2321)(0.7129)} = 26.039$$

$$n_{AO}^{(2)} = 186 \times \frac{2(0.2321)(0.7129)}{0.2321^2 + 2(0.2321)(0.7129)} = 159.9607$$

$$\mu_A^{(2)} = \frac{2n_{AA}^{(2)} + n_{AO}^{(2)} + n_{AB}}{2n} = 0.2160$$

Gene counting with the ABO blood system

Several iterations gives the following table

Iteration i	$\mu_A^{(i)}$	$\mu_B^{(i)}$	$\mu_O^{(i)}$
0	.3000	.2000	.5000
1	.2321	.0550	.7129
2	.2160	.0503	.7337
3	.2139	.0502	.7359
4	.2136	.0501	.7363
5	.2136	.0501	.7636

Disease mapping

A major topic in biostatistics and statistical genetics is disease mapping. Here the idea is that some genes are partly responsible for, or influence, the presence or absence of a disease, or maybe the severity of a disease. Sometimes you'll a phrase like "the genetic determinants of asthma", which I don't like, because it is better to say the genes are influencing the person and their traits rather than determining them. The word "disease" should be understood loosely. Historically, the main interest was in diseases, but the techniques can be applied to try to find the genes associated with any trait of interest, for example height and longetivity, as well as more traditional diseases like Alzheimer's disease, macular degeneration, and asthma.

In many cases, it is believed that rather than a single gene determining whether or not a disease occurs, it is believed that several, maybe dozens, of genetic locations each contribute to the risk and/or severity of the disease, so that individuals can either have the higher or lower risk allele at different locations in the genome. Typically, genetics will also interact with the environment, so that environmental factors, as well as combinations of genes and environment affect a person's disease status.

Disease mapping

An example of gene by environment interaction, when there is a relatively controlled environment, is if having an allele affects one's response to a drug. In this case, the drug might help relieve a certain trait (say, high blood pressure) if a person has one allele, and not help (or hurt) the person if they have the other allele. It may be that a drug helps people with blood pressure on average, but actually it might help 20% of patients who have a certain allele and have no affect on the remaining 80% of the population. When testing the null hypothesis of no effect, a large enough sample size will overcome the fact that there is no effect for the majority of people, so that there appears to be a small effect "on average".

Thinking of this type of example leads to the idea of personalized medicine, where the optimal medication for an individual might depend on their particular genes (i.e., alleles), rather than taking medicine that helps people on average, but not necessarily for an individual.

Disease mapping

For disease mapping, we can think of the human genome as having 23 chromosomes, and instead of paying attention to whole genes, like the gene for blood type, we just pay attention to positions where there are individual DNA letters.
Each gene can be thought of as a string of 100s to 1000s of DNA letters. For a given gene, most of these letters will be exactly the same for the same genomic region in two individual. If two people are chosen at random, and their DNA is lined up, then on average, about 1 in 1000 DNA letters will disagree. The entire genome for humans is about 3.3 billion letters long, and there are several million positions (roughly 0.3%) where there is some variation in the human population. In most of these cases, although there are 4 DNA letters, only 2 show up as variations within humans. If you look at more species, then you'll find more variation for different species. These single positions that have variation are called SNPs.

Disease mapping

The simplest idea for disease mapping is to code these variants as 0 or 1 (traditionally 0 is the ancestral state and 1 codes for the mutation, but it doesn't really matter). For each individual, you have 0 or 1 for some position and 0 or 1 for the presence or absence of the disease. Alternatively, if you have genotypes available, you might have say, genotypes AA, AC, CC, and you can code this gene by the number of C alleles. You can represent an individual by just these two variables (if you don't keep track of anything else): allele and disease status. Given a large sample of individuals (usually several thousand, often 10s of thousands), you can compute a simple correlation between allele and disease status.

If the allele has nothing to do with the disease status, then this correlation should be close to 0. If the correlation is high, then people with one or the other allele are more likely to have the disease.

Suppose you get a very high correlation between say, Alzheimer's in elderly patients and the number of `T` alleles at a SNP. Does this mean that the SNP causes Alzheimer's?

Disease mapping

No. It might be the case that the SNP where the association occurred is close to a gene that is causally related to the condition. In this case, since chunks of DNA get inherited together, if one mutation arose that caused the disease, it will have arisen on a single person's DNA who happened to have other genetic variants nearby. These alleles that are nearby are all closely correlated, so if one is correlated with a disease, then other nearby variants will also be related.

Over time, the association of alleles to nearby alleles decreases due to recombination. In the process of inheriting DNA, DNA gets broken into pieces at somewhat random locations (not uniformly at random). You get one chromosome from each parent. However, when you pass on DNA to your kids, your parents' DNA get shuffled and recombined. So although you may have inherited 1 copy of chromsome 7 from your mother and 1 copy from your father, your child might get 2/3 of your mother's chromosome 7 and 1/3 of your father's chromosome 7 combined to make a new chromosome 7. You child will also inherit a mixture of chromosome 7s from their other grandparents.

GWA studies

Signficant associations can be caused by

1. causal relationship between the SNP and trait
2. linkage disequilibrium between the SNP and a causal SNP (the causal SNP is nearby and the SNPs aren't independent)
3. population structure (this means the association is spurious and is misleading)
4. chance (if you are doing 100,000+ tests, some associations might appear strong

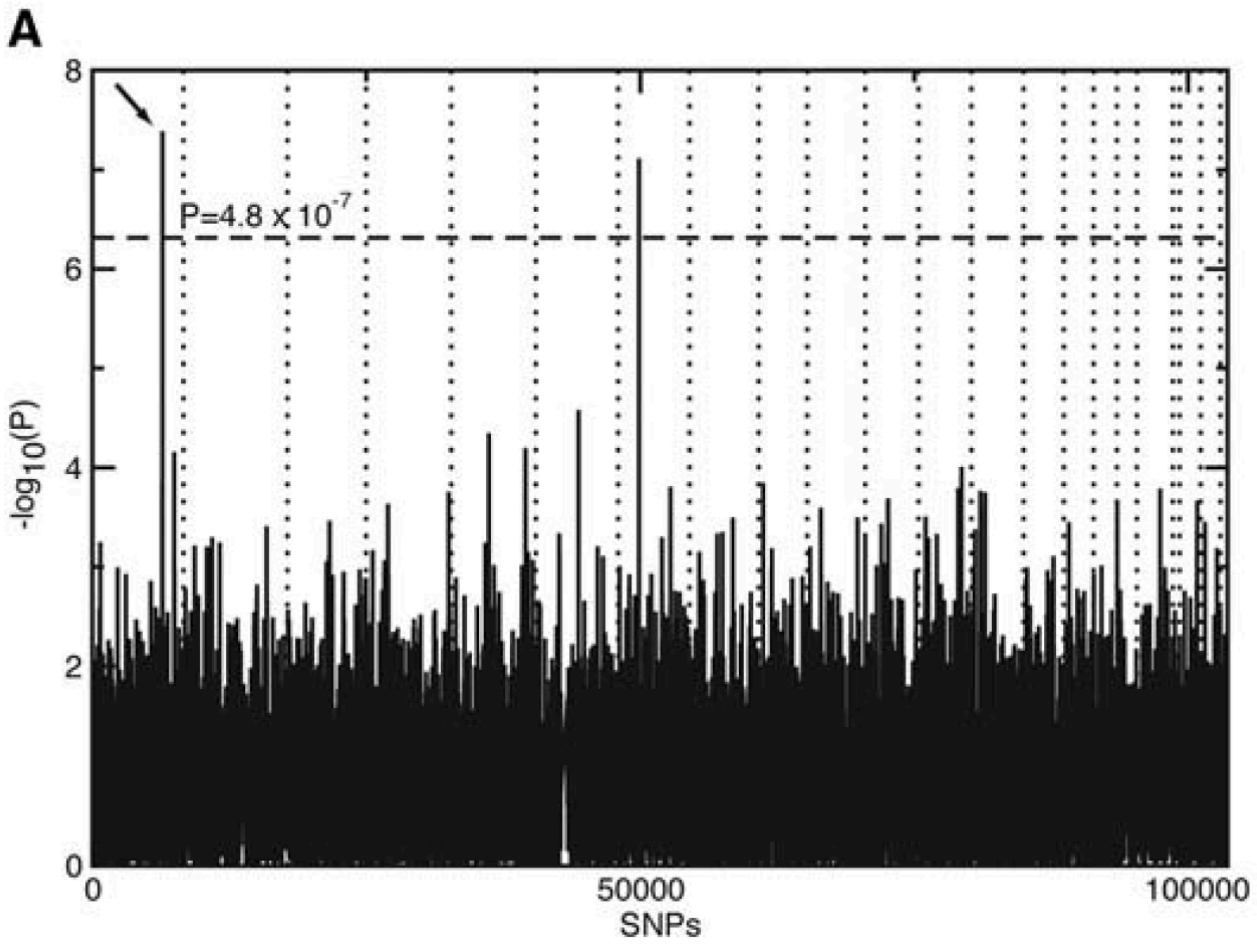
Linkage disequilibrium occurs if $P(A_1A_2) \neq P(A_1)P(A_2)$ where A_i is the probability of having allele A at gene i.

Disease mapping

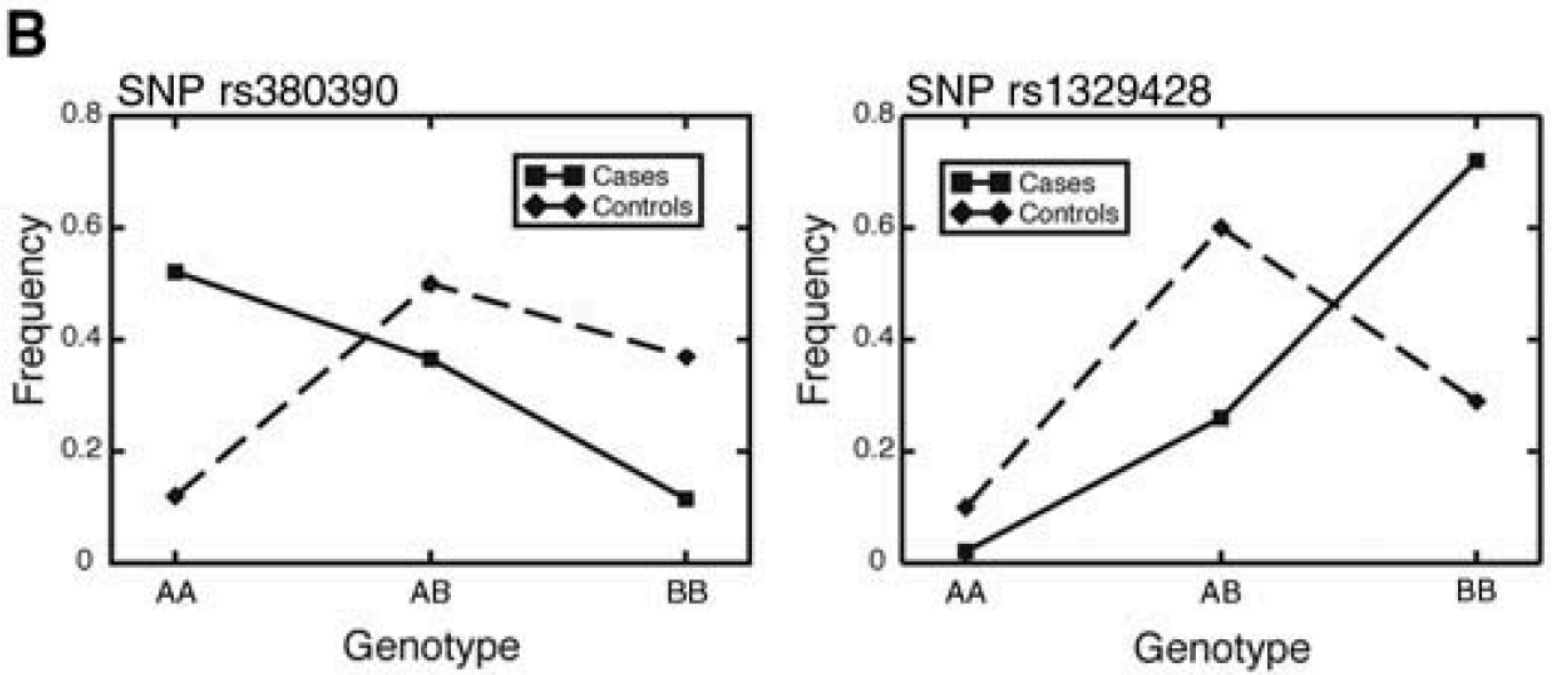
The point of this is that when you find a SNP that is highly associated with a disease, this gives you evidence that there is a SNP nearby the found SNP that influences the disease, but might not be the original SNP tested. The more SNPs you use, the more likely you are to get close to the causal SNP. Since correlation with the causal SNP should decrease with distance, stronger associations are also likely to be closer to the causal SNPs.

The usual approach then is to compute these correlations and get p-values for thousands, hundreds of thousands, or over one million SNPs. You are then computing (hundres of) thousands of p-values, one for each SNP. False positives then becomes a real problem, so your significance level needs some sort of correction. A conservative approach is to use Bonferroni adjustments.

Macular degeneration example



Macular degeneration example



Maculuar degeneration example

This study had 96 cases and 50 controls, and 116,204 SNPs genotyped, with 103,611 retained (for quality control and being autosomal). They used a Bonferroni correction of $.05/103611 = 4.826\text{e-}07$.

The study detected two significant SNPs using a Bonferroni correction, and both were located in the same intron of a gene (Complement Factor H, or CFH) on chromosome 1q31 (chromsome 1, band 31 of the long arm) which had also been found to be associated with AMD (Age-related macular degeneration) in earlier family-based studies. The study looked at the relative risk of AMD with the different combinations of SNPs (haplotypes).

GWA studies

Because the SNPs were nonfunctional, but within a gene, they looked for other SNPs in the same gene based on other studies (that weren't included in their data. This leads to hypotheses about how mutations might have functional relationships to the disease in question. Risk factors for AMD increased by a factor of 7.4 for some genetic variants found in the study. This study dated from 2005, and these days, sample sizes tend to be much larger and many more SNPs are used, but the ideas are much the same. The paper was called

"Complement Factor H Polymorphism in Age-Related Macular Degeneration", by Klein et al., Science 308(5720): 385389, 2005.

GWA studies

A problem with GWA studies has been replicability. Often a significant association is found, and then a follow-up study might not replicate the previous study or might find new variants that appear to be significant. Sometimes this is called "the winner's curse". Often if a study has low power but is lucky in finding an association or an effect, then it might be difficult to replicate that finding in future studies, so that future studies are likely to find a weaker association or weaker effect for the same SNP.

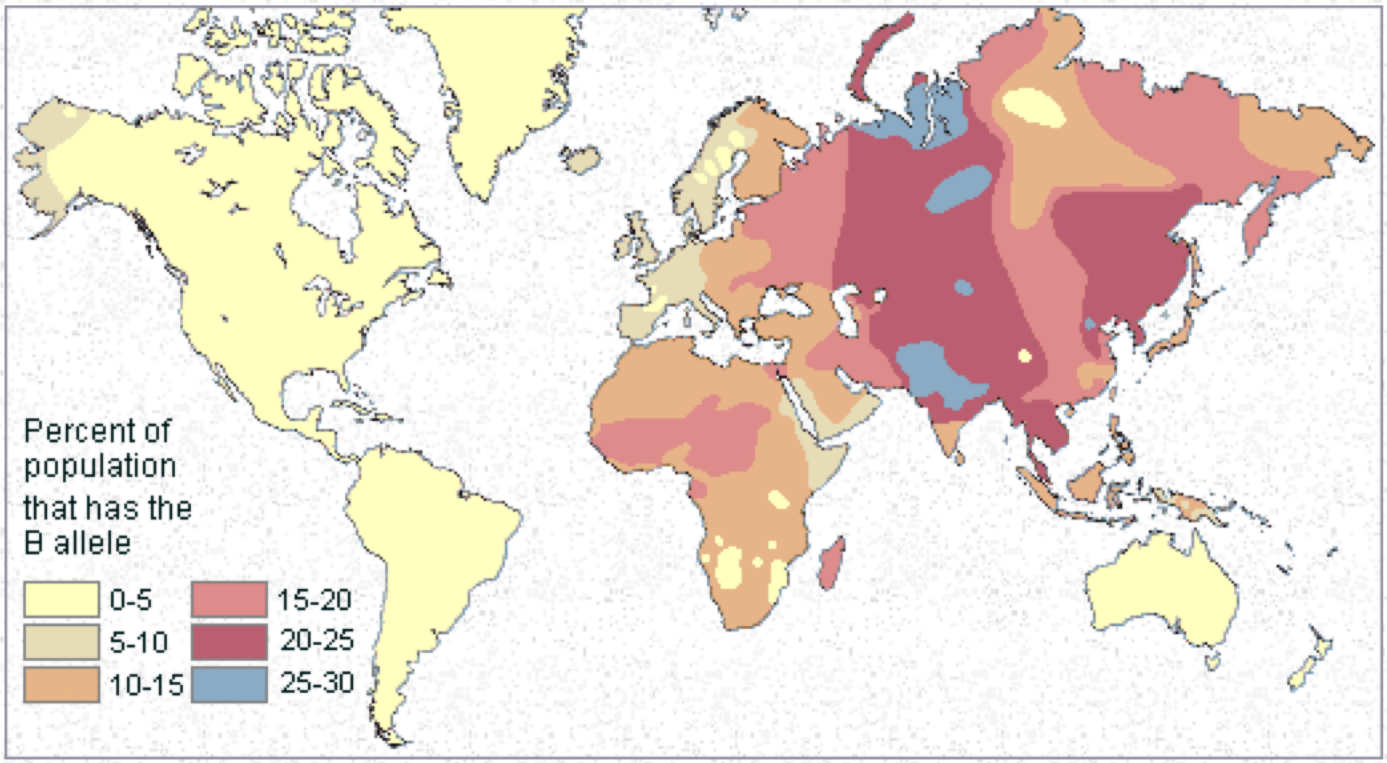
The winner's curse means that follow up studies often need to have larger sample sizes to get similar levels of significance as the original study.

GWA studies

Another problem with GWA studies is that they are sensitive to genetic structuring of populations. In particular, some diseases might be more common than others in some populations (for example, due to differences in diet, environment, health care, etc.), and allele frequencies might also differ, causing spurious correlations. One way to deal with this is to have narrowly defined populations within a GWA and to test for population structure in the data.

Another way to deal with this problem is to use family data, where you look at the distribution of allele frequency and disease status conditional on the parents genotype information. This way, the probabiity distribution of the alleles in the offspring only depends on the parents and not the population from which the offspring was sampled.

ABO allele distribution



Distribution of the B type blood allele in native populations of the world

GWA studies

To get a closer look at how this works, we can test the association between the genotype frequencies and disease status using a χ^2 test. Here "affected" means that you have the disease, whereas "unaffected" means that you don't. You can do the same thing with a continuous trait such as height or for asthma studies, the forced expiratory volume in one second (FEV1), by just doing a correlation.

	SNP genotype			
	11	12	22	total
affected	274	12	0	286
unaffected	215	26	1	242
Total	489	38	1	528

χ^2=10.83, DF 2, p=.004

GWA studies

An approach for dealing with population structure is to use family information. In this study design, you have offspring, some or all of whom are affected (when dealing with a binary trait). The original version of the test was developed in 1993 by Spielman, McGinnis and Ewens (1993) and called the Transmission-Disequilibrium Test (TDT). For the original version, the traits are binary, there is one affected offspring and two parents (called a trio design). The test has been generalized for continuous traits, multiple siblings, and missing genotype information in the parents.

GWA studies

In the original version, the idea is to look at two heterozygous parents, so they have genotypes AB. We don't worry about whether this is phenotypically distinguishable from either AA or BB.

TDT test (slides found online – no author given)

Trios: Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	w	x
	a	v	z

- w AA parents (transmit one A, do not transmit other A)
- z aa parents (transmit one a, do not transmit other a)
- x Aa parents that transmit A, do not transmit a
- y Aa parents that transmit a, do not transmit A

Possible Parental Configurations

- **AA-AA**, **AA-Aa**, **AA-aa**, Aa-AA, **Aa-Aa**, **Aa-aa**, aa-AA, aa-Aa, **aa-aa**
  - (Ones not bolded are symmetric for what we will do next, e.g., AA-Aa == Aa-AA
  - Six possible configurations

Both parents homozygous

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	2	0
	a	0	0

AA - AA
|
AA

- Offspring genotype is deterministic, no variation, not informative!

Both parents homozygous

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	0	0
	a	0	2

aa-aa
|
aa

- Offspring genotype is deterministic, no variation, not informative!

Both parents homozygous

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	1	0
	a	0	1

AA-aa
|
Aa

- Offspring genotype is deterministic, no variation, not informative!

One parent heterozygous

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	1	1
	a	0	0

AA-Aa
|
AA,Aa
.5 .5 ← Pr

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	1	0
	a	1	0

- Variation from one parent

One parent heterozygous

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	0	1
	a	0	1

Aa-aa
|
Aa,aa
.5 .5 ← Pr

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	0	0
	a	1	1

- Variation from one parent

TDT test

Both parents heterozygous

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	0	2
	a	0	0

Aa-Aa

AA,Aa,aa

.5 .5 ← Pr

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	0	1
	a	1	0

		Non-transmitted parental allele: A	Non-transmitted parental allele: a
Transmitted parental allele	A	0	0
	a	2	0

- Variation from both parents

Trios: Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	w	x
	a	v	z

- w AA parents (transmit one A, do not transmit other A)
- z aa parents (transmit one a, do not transmit other a)
- x Aa parents that transmit A, do not transmit a
- y Aa parents that transmit a, do not transmit A

Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	w	x
	a	y	z

- No variation in w or z (recall homozygous parents non informative)
- $(x-y)^2/(x+y) \sim \chi_1^2$; it's just special case of McNemar's test
- Think of it as testing are there an excess of the A allele in the affected offspring than would happen by Mendel's laws?

Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	?	78
	a	46	?

Insulin Dependent Diabetes Mellitus (IDDM)

- Example from the text: 94 families, 78 parents transmit allele A, 46 transmit allele a
- $(78\text{-}46)^2/(78+46)=8.26$, p-value=0.004

Spielman et al., 1993

FBAT

Generalizations of the original TDT test are called FBATs (Family-Based Association Studies). Compared to case-control association studies, which are more common, a trio from an FBAT design is more informative than 3 individuals in a case-control design, but it is harder to recruit trios or other family structures for a study design, and there can be a lot of missing data, not to mention incorrect data (that can't really be the father....). Consequently, the overall sample size for family-based designs tends to be smaller.

There are proponents of both approaches, but it matters which approach you want to adopt before collecting the data because the study designs are so different. The debate is one that shows up elsewhere in statistics: power versus robustness. The case-control approach tends to be more powerful because it can get larger sample sizes, but advocates for FBATs argue that population based (i.e., not family based) are less robust for population structure. Advocates for case-control designs have in turn argued that you can test for population structure and account for it...

TDT and FBATs

I don't want to take sides in the debate — I just want to point out that there has been a debate (among biostatisticians) in this area. The debate also sounds to me remarkably similar to debates about, for example, nonparametric versus parametric methods, where there is a tradeoff between power and robustness.

TDT and McNemar's test

Statistically, the TDT test is essentially McNemar's test, a test that comes from the analysis of 2×2 contingency tables.

Often in a contingency table, you find an association between say, a treatment versus a population. For example, you might have a control group given a placebo and a treatment group given a drug. The two groups are independent, and you track whether or not they experience a symptom. A famous example is an aspirin trial where doctors were randomized into placeo versus aspirin groups, and then checked for whether they had a heart attack or not within a given amount of time.

TDT and McNemar's test

For McNemar's test, instead of indepent control and treatment groups, you have correlated observations where the same individual is given both the control and the treatment at different points in time. So you might have a medication for preventing migraines, and you follow patients for 1 week with placebo and 1 week with the drug and check whether or not they experienced a migraine in the week.

	Second week	
First Week	Migraine	No migraine
Migraine	a	b
No migraine	c	d

McNemars test is

$$X^2 = \frac{(b-c)^2}{b+c}$$

TDT and McNemar's test

This has an approximate χ^2_1 distribution. Basically, this looks at whether cases in which headache status changed from week to week were different for those on the drug versus those on the placebo. If just as many people had headaches with the drug but not with the placebo as with the placebo but not with the drug, then the test statistic is 0. The diagonals do not contribute to the test statistic.